도합유사도를 이용한 한국어 문서요약 시스템

A Korean Text Summarization System Using Aggregate Similarity

김 재 훈* 김 준 홍**
(Jae-Hoon Kim) (Jun-Hong Kim)

요 약 본 논문에서 문서는 문서관계도라고 하는 가중치 그래프로 표현된다. 노드는 문서의 구성요소인 문장을 명사벡터로 표현하고, 링크는 노드들 간의 의미적인 관계를 표현하며 의미적 유사도를 가중치로 가지고 있다. 한 노드와 인접한 노드들 사이의 유사도 합을 도합유사도라고 하며, 이를 문서에서 문장의 중요도로 간주한다. 본 논문에서는 도합유사도를 이용한 한국어 문서요약 시스템을 기술한다. 실험에 사용된 평가용 요약문서는 정보처리 관련 분야에서 수집된 논문 100편과 KORDIC에서 구축한 신문기사 105건을 이용하였다. 문서요약 시스템에 의해서 생성된 요약문서의 크기가 본문의 20%이고 평가용 요약문서가 논문(서론과 결론)일 경우, 재현율과 정확률은 각각 46.6%와 76.9%를 보였으며, 또한 평가용 요약문서가 신문기사일 경우, 재현율과 정확률은 각각 30.5%과 42.3%를 보였다.

주제어 명사추출, 도합유사도, 문서요약

Abstract In this paper, a document is represented as a weighted graph called a text relationship map. In the graph, a node represents a vector of nouns in a sentence, an edge completely connects other nodes, and a weight on the edge is a value of the similarity between two nodes. The similarity is based on the word overlap between the corresponding nodes. The importance of a node, called an aggregate similarity in this paper, is defined as the sum of weights on the links connecting it to other nodes on the map. In this paper, we present a Korean text summarization system using the aggregate similarity. To evaluate our system, we used two test collections: one collection (PAPER-InCon) consists of 100 papers in the field of computer science; the other collection (NEWS) is composed of 105 articles in the newspapers and had built by KORDIC. Under the compression rate of 20%, we achieved the recall of 46.6% (PAPER-InCon) and 30.5% (NEWS), and the precision of 76.9% (PAPER-InCon) and 42.3% (NEWS).

Keywords Noun extraction; Aggregate similarity; Text summarization.

1. 서론

가상공간(cyberspace)이라고 하는 웹은 전세계를 통하여 많은 정보를 쉽게 얻을 수 있는 정보의 보고이다. 가상공간에 존재하는 정보들은 매우 다양하며, 그 양도 매우 빠른 속도로 증가하고 있다. 방대한 정보공간에서 유용한 정보를 찾기 위해 널리 사용되는 도구가 웹 정보검색 엔진이다. 일반적으로 웹 정보검색 엔진들은 너무 많은 정보를 검색해 주기 때문에 유용한 정보를 찾는 것이 그다지 쉬운 일은 아니다. 이와 같

* 한국해양대학교 컴퓨터공학과 및 첨단정보기술연구센터
주 소: 606-791 부산 영도구 동삼2동 1번지
전 화: 051-410-4574 / 팩 스: 051-404-3986
전자우편: jhoon@hanara.kmaritime.ac.kr
연구분야: 한국어정보처리, 자연언어처리, 정보검색, 정보추출
** (주)휴니드 테크놀로지스 기술연구소/통신연구팀
주 소: 435-714 경기도 군포시 당정동 352
전 화: 031-450-2763 / 팩 스: 031-459-5364
전자우편: rainmk@huneed.com
관심분야: 정보추출, 정보검색, 지능형 에이전트, 임베디드 운영체제(emgedded OS)

은 정보검색 환경에서 유용한 정보를 효과적으로 찾기 위해서는 자동문서요약 기술이 자주 사용된다[1-2].

문서요약은 원문서의 의미를 유지하면서 원문서의 길이나 정보의 복잡도를 줄이는 작업이다[2]. 즉, 문서요약은 정보압축(information compression)이다. 문서요약은 일상적인 생활에서는 널리 사용되고 있는 방법이다. 예를 들면, 뉴스의 머릿기사, 각종 회의의 의사록, 책이나 CD 등의 논평 등이 일상적인 생활 속의 문서요약에 대한 예이다. 최근 문서요약은 단순한 하나의 문서의 내용을 요약하는 것이 아니라 여러 문서의 내용은 하나로 요약하기도 하고, 심지어는 문서가 아닌 이미지, 오디오, 비디오와 같은 멀티미디어 정보를 요약하기도 한다[3].

인터넷의 급속한 발전과 더불어 문서요약에 대한 관심이 고조되면서 문서요약에 대한 연구개발에 대한 투자도 꾸준히 증가하고 있다. 특히 문서요약은 상업분야, 통신산업 분야(British Telecom의 Prosum[1]), 웹 정보검색 여과기(AltaVista Discovery에 사용되는 Inxight사의 LinguisticX[2]), 워드프로세서(Microsoft의 AutoSummarize[3]), 정보검색 색인기(National Research Council의 Extractor[4]) 등 매우 다양한 분야에서 개발되고 있다.

일반적인 문서요약 시스템은 문서분석(document analysis), 문서변환(document transformation), 문서생성(document synthesis) 단계를 거친다[5]. 문서분석은 주어진 문서를 분석해서 명사를 추출하거나 빈도수를 추출하는 등의 작업을 수행하는 단계이고, 문서변환은 분석된 정보를 토대로 본문을 요약문서로 변환하는 단계이며, 문서생성은 요약문서의 가독성을 높이기 위해서 자연스러운 문장을 생성하는 단계이다.

문서요약 기법은 크게 문장이해를 기반으로 하는 언어학적 접근 방법[5-6]과 단어 빈도수 등과 같은 통계정보를 기반으로 하는 통계적 접근 방법[7-8]으로 나눌 수 있다. 전자는 언어학적인 분석을 통하여 필요한 정보를 추출하여 이를 토대로 새로운 요약 문서를 생성한다. 이 방법은 문서의 이해가 어려울 뿐 아니라 많은 언어적 지식이 요구된다. 후자는 문서 중에서 중요한 문장 혹은 단락을 추출하고 중요도에 따라 재배치하는 방법으로 문서를 요약한다. 후자에 의해 생성된 요약문서는 전자의 것에 비해 가독성이 떨어지며 최근에 가독성을 높이기 위한 연구들이 막 진행되고 있다[9-11].

최근에 와서 한국어 문서요약에 대한 연구도 매우 활발히 진행되고 있다[2,12-13]. 그러나 아직은 성숙되지 않은 것 같다. 또한 대부분의 연구가 통계적 접근 방법을 채택하고 있으며, 여러 다양한 환경에서 평가되어 객관적으로 어떤 시스템이 좋은 성능을 보인다고 말할 수 없는 실정이다.

문장을 기본 요소로 하여 하나의 문서를 그래프로 표현할 수 있으며, 이를 문서관계도(text relationship map)라고 한다[9]. 문서관계도에서 노드는 문장 혹은 단락을 표현하고, 링크는 의미적으로 관련된 노드들 사이의 관계를 나타낸다. 이 관계는 노드와 노드 사이의 유사도가 어떤 임계값 이상일 경우를 말한다. 무성도(bushiness)는 문서관계도에서 다른 노드와 연결된 링크 수(일명 부쉬경로(bushy path)), 즉 노드의 차수이며, 무성도가 높으면 높을수록 많은 다른 노드들과 많이 연결되었음을 의미한다. 문서요약은 단락이나 문장을 무성도가 높은 순으로 재배치하는 것이다. 본 논문에서는 문서관계도에서 무성도 개념[9]을 이용하였다. 무성도는 임계값에 의해서 문장들 사이의 유사관계가 정의된다. 즉, 문장의 유사도가 임계값 이상인 문장들은 유사한 문장이라고 판단하고 유사한 문장이 많을수록 중요한 문장으로 간주한다. 따라서 무성도를 이용할 경우, 문서의 종류가 바뀌거나 영역이나 장르 등이 변할 경우 다른 임계값을 사용한다. 이런 문제를 다소 완화하기 위해서 본 논문에서는 무성도를 단순화한 도합유사도(aggregate similarity)를 정의하였으며, 도합유사도는 유사도의 합이다.

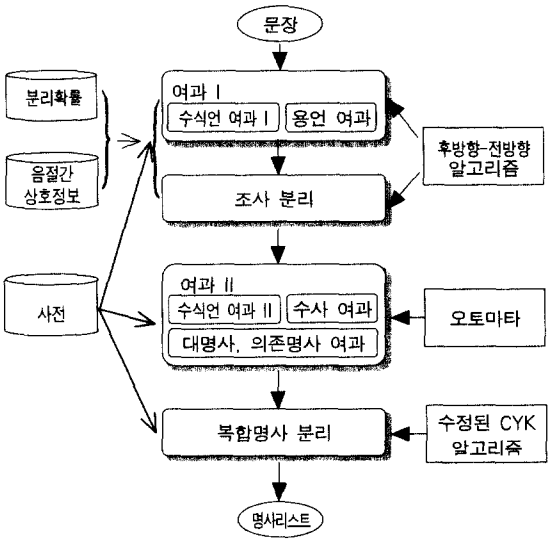
본 논문의 구성은 다음과 같다. 2장에서 한국어 명사 추출 시스템에 대해서 기술하고, 3장에서는 도합유사도에 대해서 자세히 기술한다. 4장에서는 한국어 문서 요약 시스템을 설명한다. 5장에서 실험 및 평가에 대해서 기술하고, 6장에서 기존의 제안된 방법들과 비교하고 분석한다. 마지막으로 7장에서 결론을 맺고 앞으로의 연구 방향에 대해서 논의한다.

2. 한국어 명사 추출

본 논문에서 문장은 벡터공간의 한 점으로 표현되고 이를 문장벡터라고 한다. 문장벡터는 각 명사의 빈도수에 의해서 표현된다. 따라서 문장벡터를 구성하기 위해서는 문장에 포함된 명사를 추출해야 한다. 한국어 명사추출에 관한 많은 연구는 한국어 정보검색 분야에서 많이 수행되었다[14-16]. 본 논문에서는 여과

1) http://transend.labs.bt.com
2) http://www.inxight.com
3) http://microsoft.com
4) http://extractor.iit.nrc.ca/

기법과 분리기법을 이용한 한국어 명사추출 시스템[17]의 구조는 (그림 1)과 같으며, 다음과 같은 절차에 의해서 명사를 추출한다.



(그림 1) 한국어 기준명사 추출 시스템의 구조

여과 단계 I에서는 명사를 포함하지 않는 어절을 사전과 후방향-전방향 알고리즘을 이용하여 여과하는 단계이며, 후방향-전방향 알고리즘에서는 두 음절 사이에서 어미가 분리될 확률과 두 음절이 어미에 속할 상호정보(mutual information)[18]를 이용한다. 수식언(부사, 관형사)과 독립언(감탄사)을 여과하기 위해서는 사전을 이용하고, 용언(형용사, 동사)을 여과하기 위해서는 후방향-전방향 알고리즘을 이용한다.

명사가 포함된 어절은 주로 명사[5]와 조사[6]로 구성되는데 정확한 명사를 추출하기 위해서는 조사가 분리되어야 한다. 조사 분리 단계에서 후방향-전방향 알고리즘을 이용하여 조사를 분리한다[17]. 여기서 분리확률은 두 음절 사이에서 조사가 분리될 확률이고, 상호정보는 두 음절이 조사에 속할 상호정보이다.

여과 단계 II에서도 명사를 포함하지 않는 어절을 여과한다. 부사는 일반적으로 보조사와 결합이 가능하다. 따라서 조사를 분리한 후에 사전을 이용해서 수식언에 속하는 부사를 제거한다. 또한 이 단계에서는 명사라고 하더라도 문서요약에서 필요로 하지 않는 수사, 대명사, 의존명사를 여과한다. 수사는 오토마타를 이용해서 여과하고, 대명사와 의존명사는 사전을 이용해서 여과한다.

문서요약의 강인성을 극대화하기 위해서 본 논문에서는 복합명사를 분리한다. 정확률을 높이기 위해서는 복합명사를 분리하지 않는 것이 좋을 것이다. 복합명사분리는 사전과 수정된 CYK 파싱 알고리즘[17, 19]을 이용한다.

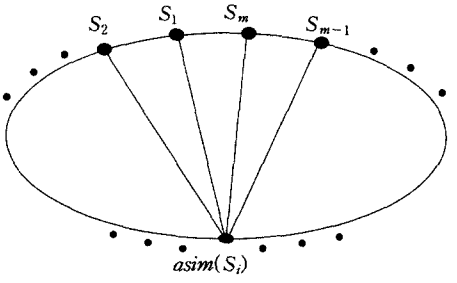
3. 도합유사도

문장을 기본 요소로 하여 하나의 문서는 문서관계도[9]로 표현될 수 있으며, 문서관계도는 일종의 그래프이며, 본 논문에서는 완전그래프이다. 문서관계도에서 노드는 문장벡터를 나타내고, 링크는 의미적으로 관련이 있는 노드(문장벡터)들 사이의 관계를 나타내며, 의미적 관계 정도를 나타내는 가중치를 가진다. 의미적 관계 정도는 의미유사도이며, 두 문장벡터 S_i와 S_j 사이의 의미유사도 $\mathrm{sim}(S_i, S_j)$는 식 (1)과 같다.

$$\mathrm{sim}(S_i, S_j) = \sum_{k=1}^{n} s_{i,k} s_{j,k} \qquad (1)$$

여기서 문서에 포함된 명사의 수는 n개라고 할 때, 문장벡터 S_i는 $(s_{i,1}, s_{i,2}, \ldots, s_{i,n})$이고, $s_{i,k}$는 i번째 문장에 포함된 명사 k에 대한 문서 내의 빈도수이다. 문장의 중요도는 각 노드에 인접한 노드들과의 의미유사도의 합으로 정의되며, 이를 본 논문에서는 도합유사도라고 하며, i번째 문장에 대한 도합유사도 $\mathrm{asim}(S_i)$는 식 (2)와 같이 정의되며, (그림 2)는 m개의 문장으로 구성된 문서에서 i번째 문장에 대한 도합유사도를 개념적으로 도식화한 것이다.

$$\mathrm{asim}(S_i) = \sum_{\substack{j=1 \\ j \neq i}}^{m} \mathrm{sim}(S_i, S_j) \qquad (2)$$

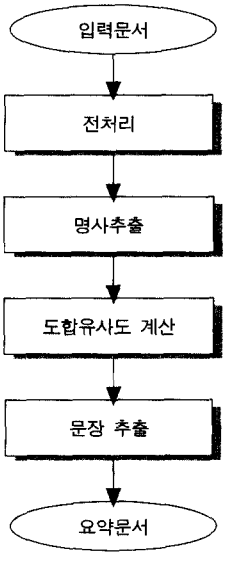


(그림 2) 도합유사도의 개념도

5) 복합명사도 포함된다.
6) 조사는 생략될 수도 있고 복합조사도 포함된다.

4. 한국어 문서요약 시스템

본 논문의 문서요약은 추출요약(indicative summarization)에 해당되며, 통계적인 접근 방법을 사용한다. 일반적으로 추출요약은 문장의 중요도를 계산하고 그 중요도를 기준으로 요약문서에 포함될 문장을 추출한다. 본 논문에서는 문장은 명사목록으로 표현되며, 문장 간의 유사도는 내적(inner product)을 사용하며, 문장의 중요도는 도합유사도를 이용한다. 문장생성은 문장의 중요도가 높은 순으로 정렬한다. 또한 본 논문은 특별한 학습기법을 전혀 사용하지 않는다. 이와 같은 개념을 토대로 본 논문에서 제시한 한국어 문서요약 시스템의 구조는 (그림 3)과 같다.



(그림 3) 한국어 문서요약 시스템의 흐름도

전처리기는 입력문서를 문장 단위로 분리하고, 문장 기호를 제거한다. 문장을 분리하는 기준은 기호 ". ! ?"가 있으면 문장으로 분리하였다. 이외에 "1999. 12." 등과 같은 문자열에 대해서 오류가 발생하기 때문에 약간의 경험규칙(heuristics)을 사용하여 해결한다.

명사추출 방법과 도합유사도의 계산 방법은 각각 2장과 3장에서 구체적으로 설명하였다.

문장추출은 먼저 도합유사도가 높은 순으로 문장을 재정렬하고, 그리고 나서 도합유사도가 높은 순으로 원하는 비율만큼의 문장을 추출하여 요약문서로 출력한다.

5. 실험 및 평가

5.1. 실험 말뭉치

본 논문에서는 평가를 위해 두 종류의 평가용 말뭉치를 구축하였다. 하나는 정보처리 관련 분야의 논문 100편(PAPER)으로 구성되었다. 이 말뭉치의 구축 방법은 Teufel과 Moens이 제안된 방법[20]과 비슷하게 원저자의 초록(abstract)을 이용하였다. 즉, 논문의 전체(PAPER-ALL) 혹은 서론과 결론 부분(PAPER-InCon)에 속하는 문장들 중에서 원저자의 초록에 속하는 문장에 가장 유사한 문장들을 추출하여 평가용 말뭉치를 구축하였다. 이때 평가요약문서의 크기는 원저자의 초록의 크기와 같고, 유사도로 내적을 사용하였다. 또 다른 하나는 KORDIC 말뭉치의 신문기사 105건(NEWS)이고, 이 말뭉치의 평가용 요약문서는 요약전문가에 의해서 구축되었다[16]. 이들 두 말뭉치의 통계치는 〈표 1〉과 같다. PAPER-InCon과 NEWS의 통계치들은 서로 비슷하며, 장르는 서로 다르다.

〈표 1〉 평가용 말뭉치의 통계치

	100	100	105
	113.2	20.6	21.5
	5.6	5.6	6.0

5.2. 성능 평가

성능평가의 측도는 정보검색 분야에서 널리 사용되고 있는 정확률(precision)과 재현율(recall) 그리고 F측도(f-measure)를 사용하였으며, 이들은 각각 식 (3 ~ 5)과 같이 정의된다[21].

$$P = \frac{N_R}{N_S} \quad (3)$$

$$R = \frac{N_R}{N_C} \quad (4)$$

$$F = \frac{2PR}{P+R} \quad (5)$$

여기서 N_S는 문서요약 시스템이 제시한 전체 문장 수이고, N_R은 N_S 중에서 평가 요약문서에 속한 문장 수이고, N_C는 평가 요약문서에 속한 문장 수이다. 〈표 2〉는 제안된 시스템의 정확률(P), 재현율(R), F측도(F)를 보이고 있으며, 본 논문에서는 요약문서의 크기가 원문의 10%와 20%에 대한 성능을 보일 것이다[7].

〈표 2〉 제안된 시스템의 성능 평가

10%		34.3	74.2	46.9
		85.2	33.8	44.6
		46.8	13.3	20.7
20%		19.7	83.3	31.9
		76.9	46.6	58.0
		42.3	30.5	35.4

PAPER-ALL의 본문은 다른 말뭉치에 비해서 크므로 다른 말뭉치와는 반대로 재현율이 높은 결과를 가져왔다. 일반적으로 논문에 대한 문서요약을 할 경우 본문을 서론과 결론만 이용하는데 이를 실험적으로 관찰해보고자 PAPER-ALL을 말뭉치로 채택하였으며, 일반적인 서론과 결론을 이용하는 방법의 어느 정도는 타당함을 알 수 있었다. PAPER-InCon과 NEWS는 서로 비슷한 말뭉치의 특성을 가지고 있으나 성능은 크게 차이를 보인다. 이 원인은 크게 두 가지로 요약될 수 있다. 하나는 평가 요약문서의 구축 방법의 차이이다. PAPER-InCon은 자동으로 구축되었고, NEWS는 요약전문가에 의해서 구축되었다. 자동으로 구축된 PAPER-InCon의 평가용 요약문서는 문서요약 시스템의 특성이 일부 반영되었다고 볼 수 있다. 다른 하나는 장르의 차이이다. 기술논문에는 반복적인 표현들이 자주 발생된다는 특징을 잘 반영하는 것으로 추정된다. 좀더 정확한 원인 분석을 위해 좀 더 많은 연구가 필요하다.

5.3. 왜 내적 유사도를 사용하는가?

본 논문에서 문장 간의 유사도로 내적을 사용하였다. 내적 유사도를 사용한 특별한 이유가 있는 것은 아니다. 일반적으로 정보검색 시스템에서 유사도 계산을 위해서 코사인이 널리 사용되는데 본 논문에서는 성능 평가를 통해서 내적 유사도를 사용하게 되었다. 〈표 3〉은 내적 유사도와 코사인 유사도의 성능을 비교한 것이다. 내적 유사도는 코사인 유사도에 비해 계산이 용이할 뿐 아니라 문장의 길이에 대한 특성을 반영하고 있다.

〈표 3〉 내적 유사도와 코사인 유사도의 비교

10%	PAPER-InCon	85.2	33.8	44.6
		77.6	22.3	34.6
	NEWS	46.8	13.3	20.7
		31.9	8.7	13.7
20%	PAPER-InCon	76.9	46.6	58.0
		28.6	50.0	36.4
	NEWS	42.3	30.5	35.4
		39.9	22.8	29.0

5.4. 왜 용언을 사용하지 않는가?

용언이 문장의 의미를 결정하는 주성분이므로 문서요약에서도 용언을 충분히 사용한다면 좋은 결과를 가져올 것으로 기대된다. 따라서 본 절에서는 용언의 문장벡터에 포함할 경우 본 논문에서 제안한 문서요약 시스템의 성능을 관찰해 보았다. 〈표 4〉는 용언을 문장벡터에 포함할 경우와 그렇지 않을 경우의 성능을 비교한 것이다.

〈표 4〉 문장벡터에 용언이 추가되었을 때, 성능 변화

10%	PAPER-InCon	85.2	33.8	44.6
		41.8	11.5	18.0
	NEWS	46.8	13.3	20.7
		25.6	7.2	10.9
20%	PAPER-InCon	76.9	46.6	58.0
		39.0	20.2	27.3
	NEWS	42.3	30.5	35.4
		27.4	15.5	19.8

〈표 4〉의 결과에서 용언이 추가되었을 때, 모든 성능이 좋지 않았다. 용언은 구문적인 성질을 가지고 있는데, 이 성질을 전혀 반영하지 않은 상태에서 문장을

7) 일반적인 요약문서의 크기는 원문서의 1% ~ 30%에 달한다[4]. 그러나, 추출요약의 경우에는 약 20%정도는 되어야 원문서의 의미를 전달할 수 있다[22].

단순한 단어의 목록으로 표현하였기 때문에 용언이 추가되어도 좋은 결과를 가져오지 못했다. 따라서 용언을 이용할 경우에는 반드시 구문분석이 필요하며, 분석된 구문구조를 이용하여 문장을 표현하여 문장의 유사도를 계산해야 한다는 것을 알 수 있었다.

6. 관련 연구와의 비교

6.1. 도합유사도와 부쉬경로

Salton은 문장의 중요도를 부쉬경로로 측정한다[9]. 1장에서 언급했듯이 부쉬경로는 문서관계도에서 링크수를 나타내며, 문장 간의 유사도가 어떤 특정 임계값 이상일 때, 링크가 설정된다. 〈표 5〉와 〈표 6〉은 각각 생성된 요약문의 크기가 본문의 10%와 20%일 때, 임계값의 변화에 따른 부쉬경로 방법의 재현율과 정확률을 보인 것이다. NEWS 말뭉치에서 도합유사도와 부쉬경로는 비슷한 성능(재현율과 정확률 모두)을 보였으며, PAPER-InCon에 대해서는 도합유사도가 훨씬 더 좋은 성능을 보이고 있다. 도합유사도는 부쉬경로의 임계값과 같은 매개변수를 가지고 있지 않기 때문에 새로운 환경에 쉽게 적용될 수 있을 것이다. 즉, 새로운 매개변수를 찾는 노력이 없이도 새로운 환경에 적응할 수 있을 것이다.

〈표 5〉 부쉬경로와 도합유사도의 성능 비교 (10%)

말뭉치	부쉬 경로			제안된 시스템	
	R	P	임계값	R	P
PAPER-InCon	24.9	65.5	0.12	33.8	85.2
	28.6	74.0	0.14		
	29.0	74.8	0.16		
	28.2	72.6	0.18		
	28.9	74.4	0.20		
	29.9	76.8	0.22		
	29.6	75.9	0.24		
	29.6	76.1	0.26		
News	13.0	48.3	0.12	13.3	46.8
	12.7	46.2	0.14		
	14.5	52.7	0.16		
	11.9	44.6	0.18		
	12.1	44.9	0.20		
	12.5	46.9	0.22		
	13.4	50.6	0.24		
	13.1	50.1	0.26		

6.2. 제안된 시스템과 상용시스템

독립적으로 상용되는 한국어 문서요약 시스템은 없으나, 워드프로세서(Microsoft Word와 훈민정음) 내의 자동요약 도구를 가지고 있다. 객관적인 성능을 비교하기 위해서 본 논문에서는 이들 상용시스템과의 성능을 평가해 보았다. 〈표 7〉는 제안된 시스템과 상용시스템과의 성능을 비교한 결과이다. 제안된 시스템이 재현율과 정확률 모든 면에서 좋은 결과를 보였다.

〈표 6〉 부쉬경로와 도합유사도의 성능 비교(20%)

말뭉치	부쉬 패스			제안된 시스템	
	R	P	임계값	R	P
Paper-InCon	37.5	62.9	0.12	46.6	76.9
	39.8	66.1	0.14		
	39.8	66.2	0.16		
	38.6	64.5	0.18		
	41.6	69.7	0.20		
	41.1	67.3	0.22		
	42.8	70.7	0.24		
	42.7	70.9	0.26		
News	22.5	39.4	0.12	30.5	42.3
	24.5	43.3	0.14		
	26.0	45.9	0.16		
	26.2	45.8	0.18		
	24.6	44.4	0.20		
	24.5	44.5	0.22		
	25.0	45.0	0.24		
	25.0	45.0	0.26		

〈표 7〉 제안된 시스템과 상용시스템과의 성능 비교

요약문의 크기	말뭉치	시스템	P	R	F
10%	PAPER-InCon	제안된 시스템	85.2	33.8	44.6
		Microsoft Word	42.1	18.8	20.8
		훈민정음	66.3	31.0	42.2
	NEWS	제안된 시스템	46.8	13.3	20.7
		Microsoft Word	29.3	12.5	17.5
		훈민정음	20.5	11.3	14.6
20%	PAPER-InCon	제안된 시스템	76.9	46.6	58.0
		Microsoft Word	31.0	24.9	27.6
		훈민정음	60.3	47.4	53.1
	NEWS	제안된 시스템	42.3	30.5	35.4
		Microsoft Word	31.5	24.7	27.7
		훈민정음	16.4	14.1	15.2

6.3. 기존의 한국어 문서요약 시스템

본 절에서는 제안된 문서요약 시스템과 기존의 한국어 문서요약 시스템들과 개념적으로 비교 분석하고자 한다. 각 시스템들은 매우 다양한 실험 환경 때문에 객관적인 성능을 비교하는 것은 어려운 일이다. 〈표 8〉은 기존 시스템들의 주요 특징을 나타낸 것이다. 대부분의 방법은 통계적인 방법을 이용하고 있으며, 기술문서의 일종인 논문을 중심으로 평가했으며, 평가용 요약문서의 수가 20~30으로 비교적 작은 규모로 실험하였다. 본 논문은 객관성을 높이기 위해서 약 100여 개의 문서에 대해서 논문뿐 아니라 신문기사에 대한 실험을 수행하였다.

또한 대부분의 시스템들이 학습모델을 이용하고 있다. 학습모델을 사용할 경우에는 학습되지 않은 영역에 적응하기 위해서는 경우에 따라서 새로운 학습이 필요하다. 본 논문에서 제안한 방법은 학습모델을 사용하지 않으며 주어진 문장에 대한 명사의 빈도수만 필요하다.

본 논문에서 제안한 한국어 문서요약 시스템은 기존의 다른 시스템에 비해서 모델이 단순하며 쉽게 구현할 수 있고 실용적으로 사용할 수 있도록 설계 구현되었다. 그렇지만, 성능면에서도 기존의 다른 시스템보다 떨어지지 않았으며, 객관적인 성능 비교는 어려울지 모르지만 재현율과 정확률에 대해서 조금 더 좋은 성능을 보였다.

7. 결론

본 논문은 도합유사도를 이용한 문서요약 시스템을 제안하였다. 한 문장의 도합유사도는 본문 내에 있는 다른 문장들과의 유사도 합을 의미하며, 유사도로 내적을 사용하였다. 본 논문의 문서요약 방법은 단순한 모델을 사용하고 있으며, 구현이 용이하고, 쉽게 실용적으로 사용할 수 있다는 장점을 가지고 있다.

제안된 시스템을 평가하기 위해 두 종류의 말뭉치(논문, 신문기사)를 사용하였다. 시스템이 생성한 요약문의 크기가 본문 크기의 20%이고, 본문이 논문(서론과 결론)일 경우, 재현율과 정확률은 각각 46.6%와 76.9%를 보였으며, 또한 본문이 신문기사일 경우, 재현율과 정확률은 각각 30.5%과 42.3%를 보였다. 또한 제안된 방법은 상용시스템보다 좋은 성능을 보였다.

통계적인 접근 방법을 이용하는 대부분의 시스템에서 생성된 요약문서는 가독성이 좋지 않다. 이를 위해서는 문서계획과 문장생성에 관한 연구가 필요로 하다. 문서요약 기술은 정보검색의 색인이나 문서분류 등의 분야에 적용하여 더욱 더 질 좋은 문서검색 시스템을 구현할 수 있을 것이다.

〈표 8〉 한국어 문서요약 시스템들의 특징

시스템	접근방법	실험환경	성능		비교 (요약문의 크기)
			P	R	
[12]	휴리스틱 방법 통계적 방법 학습함. 문장 유사도 문장단위로 추출	논문(서론과 결론) 평가 문서 수: 20	66.80		본문 크기의 30%
[13]	통계적 방법 부쉬경로 코사인 유사도 단락 단위로 추출. 가중치: tf * idf	논문(서론과 결론) 평가 문서 수: 25		35.0	본문 크기의 30%
[23]	통계적 방법 학습함. 문장 유사도 문장단위로 추출	논문(서론과 결론) 평가 문서 수: 25	51.08	42.4	본문 크기의 20%
[24]	휴리스틱 방법 통계적 방법 학습함 문장단위로 추출	논문(서론과 결론) 평가 문서 수: 30	53.19	39.53	5개의 문장
본 논문	도합 유사도 문장 단위 출력	논문(서론과 결론) 평가 문서 수:100	76.9	46.6	본문 크기의 20%
		신문기사 평가 문서 수:105	42.3	30.5	

감사의 글

본 연구는 첨단정보기술 연구센터를 통하여 과학재단과 지원을 받았으며, 또한 과학기술부 STEP2000 프로젝트에 의해 지원되고, 전문용어언어공학연구센터에 의해 수행중인 "대용량 국어정보 심층처리 및 품질관리 기술개발" 연구과제의 일환으로 수행되었습니다.

참고문헌

[1] Cowie, J., Mahesh, K., Nirenburg, S. and Zajac, R. "MINDS - Multilingual interactive document summarization." *Working Notes of the AAAI Spring Symposium on Intelligent Text Summarization*, Spring, pp. 131-132, 1998.

[2] 장동현, 맹성현, "자동 요약 시스템," *정보과학회지*, 제15권, 제10호, pp. 42-49, 1997.

[3] Mani, I. and Maybury, M. T., *Advanced in Automatic Text Summarization*, The MIT Press, 1999.

[4] Sparck Jones, K. "Automatic summarizing: factors and directions," in Mani, I. and Maybury, M. T., editors, *Advances in Automatic Text Summarization*, pp. 1-12. The MIT Press, 1999.

[5] Rau, L. F., Jacobs, P. S., and Zernik, U., "Information extraction and text summarization using linguistic knowledge acquisition," Information Processing and Management, vol. 25, no. 4, pp. 419-428, 1989.

[6] Reimer, U. and Hahn, U., "Text condensation as knowledge base abstraction," *Proceedings of IEEE Conference on AI Applications*, pp. 338-344, 1988.

[7] Kupiec, J., Pedersen, J., and Chen, F. "A trainable document summarizer," *Proceedings of the 8th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 68-73, 1995.

[8] Hovy, E. and Lin, C.-Y. "Automated text summarization in SUMMARIST," *Proceedings of the TIPSTER Workshop*, 1998.

[9] Salton, G., Singhal, A., Mitra, M. and Buckly, C. "Automatic text structuring and summarization," *Information Processing and Management*, vol. 33, no. 2, pp. 193-207, 1997.

[10] Daniel Marcu. *The rhetorical parsing, summarization, and generation of natural language texts*. PhD thesis, Department of Computer Science, University of Toronto, Forthcoming 1997.

[11] Azzam, S., Humphreys, K., and Gaizauskas, R., "Using coreference chains for text summarization," *Proceedings of the ACL'99 Workshop on Coreference and its Applications*, Maryland, 1999.

[12] 이문희, 박미성, 김미진, 이상조, "문서 특성과 제목을 이용한 문장 추출," *한국정보과학회 가을 학술발표논문집(B)*, 제26권, 제2호, pp. 441-445, 1999.

[13] 류동원, 이종혁, "단어공기정보를 이용한 자동화 문서 요약," *한국정보과학회 봄 학술발표논문집*(B), 제27권, 제1호, pp. 345-347, 2000.

[14] 박영찬, 최기선, "통계적 정보를 이용한 복합명사 검색 모델," *인지과학*, 제6권, 제3호, 1995.

[15] 채영숙, 권혁철, "말뭉치로부터 추출된 통계 정보를 활용한 한국어 복합명사 분석," *인지과학*, 제8권, 제2호, pp. 101-108, 1997.

[16] 윤보현, 조민정, 임해창, "통계정보와 선호 규칙을 이용한 한국어 복합 명사의 분해," *정보과학회논문지*(B), 제24권, 제8호, pp. 900-909, 1997.

[17] 김재훈, 김준홍, 박호진, "여과 및 분리 기법을 이용한 한국어 기준 명사 추출," *제12회 한글 및 한국어 정보처리 학술대회 발표논문집*, 성공회대학교, 서울, pp. 3-10, 2000.

[18] Church, W. K., "Word association norms, mutual information and lexicography," Computational Linguistics, vol. 16, no. 1, pp. 22-29, 1990.

[19] Aho. V. A. and Ullman, J. D., *The Theory of Parsing, Translation, and Compiling*, Prentice-Hall, 1973.

[20] Teufel, S. and Moens, M. (1999). "Argumentative classification of extracted sentences as a first step towards flexible abstracting," in Mani, I. and Maybury, M. T., editors, *Advances in Automatic Text Summarization*, pp. 155-171. The MIT Press.

[21] Manning, C. D. and Shutze, H., *Foundation Statistical Natural Language Processing*, The MIT Press, 1999.

[22] Morris, A. H., Kasper, G. M., Adams, D. A., "The effects and limitations of automated text condensing on reading comprehension performance," *Information Systems Research*, pp. 17-35, March 1992.

[23] 강상배, *한국어 문서의 통계적 정보를 이용한 문서 요약 시스템 구현*, 부산대학교, 컴퓨터공학과, 석사학위 논문, 1997.

[24] Myaeng, S. H. and Jang, D., "Development and evaluation of a statistically based document summarization system." in Mani, I. and Maybury, M. T., editors, *Advances in Automatic Text Summarization*, pp. 61-70, The MIT Press, 1999.